The 1880 U.S. Population Database

RON GOEKEN
CUONG NGUYEN
STEVEN RUGGLES
WALTER SARGENT
Minnesota Population Center
University of Minnesota

Abstract. The 1880 U.S. population database contains records for the over 50 million individuals enumerated in the census. This unique data set is the result of collaboration between the Minnesota Population Center (MPC) and the Church of Jesus Christ of Latter-day Saints (LDS). To make census data available to genealogists, LDS volunteers transcribed the characteristics of all U.S. residents in 1880. The MPC verified and corrected this transcription, in exchange for the right to disseminate the resulting database for scholarly and educational purposes. The authors consider editing and coding procedures and a range of problems: missing and incorrect geographic identifiers, data-processing errors, duplicate and missing cases, and inaccurate breaks between households. They also classify variables according to standardized coding systems, making the database compatible with the Integrated Public Use Microdata Series (IPUMS).

Keywords: 1880 census, IPUMS, microdata, population

Between 1982 and 1999, volunteers from the Church of Jesus Christ of Latter-day Saints (LDS) invested some 11 million hours transcribing the characteristics of all 50 million persons residing in the United States in 1880. The goal of the project was to create an electronic look-up system for genealogical research. Although the manuscript data are in the public domain, the machine-readable version was copyrighted by the LDS, which planned to make the data available exclusively in a proprietary format designed for genealogical searches of individual cases.

In May 1999, the Minnesota Population Center (MPC) approached the LDS about creating a version of the database suitable for demographic research. At first, the LDS responded unenthusiastically. Concerned solely with creating a high-quality genealogical product as quickly as possible, the LDS was struggling with limited technical resources to fix a number of serious data problems. When the MPC offered to clean the data in exchange for permission to distribute the database freely for academic use, the LDS readily agreed.

In fall 1999, the MPC obtained funding from the National Science Foundation to improve the data and completed basic cleaning in January 2001. The LDS released the cleaned data on a set of 55 CD-ROMs in June 2001 and unveiled a free national online look-up system for the data in October 2002. This search system constitutes an invaluable source for historical research, but it is not suitable for quantitative analysis. With funding from the National Institutes of Health, the MPC has been working since 2001 to create a numerically coded sequential version of the database that is useful for social science research. A preliminary version of this database is now available.[1]

Applications of the Complete-Count Database

Despite the high quality of the 1880 enumeration, the published tabulations from that census are of limited value. Partly because of the increased expenditure in the enumeration phase, the census appropriations were exhausted in the first half of 1881, long before the tabulations were complete. Until January 1882, when Congress reluctantly provided additional funds, the count was carried out by a skeleton staff of volunteer clerks. Another interruption resulted when Francis Walker resigned his position as superintendent of the census. When funding for the 1880 census expired in early 1885, the work was still incomplete. In the end, many of the most interesting variables collected—such as marital status, relationship to head, birthplace of mother, and birthplace of father—were never tabulated at all, and there were few cross-tabulations of any variables. No data on family characteristics were published beyond the number of families and their average size (U.S. Census Office 1882, 1883, 1895; Walker 1888; Wright and Hunt 1900; Scott 1968; Eckler 1972; Magnuson 1995).

The 1880 census falls in the middle of a major gap in published census tabulations. County tabulations of population by age, sex, race, and nativity ended in 1860; tabulations of these characteristics by locality did not begin

until 1890. The 1880 database allows us to extend the early-nineteenth-century series forward and the twentieth-century series backward. In addition to these tabulations, the database permits augmentation of other county- and city-level statistics (such as those contained in "Historical, Demographic, Economic, and Social Data: The United States, 1790–1970" from the Inter-university Consortium for Political and Social Research), to make the 1880 data more compatible with that available for other census years.

Such tabulations will greatly improve demographic estimates for the late nineteenth century. They permit far more detailed study of fertility than is presently possible, allowing researchers to carry out analyses by city and county (cf. Leet 1977; Schapiro 1985). Existing aggregate statistics do not even allow analysis of rural-urban fertility differentials at the state level in 1880 (see U.S. Census Bureau 1975). Equally important, new tabulations will advance our understanding of mortality in a key period of transition. There are vital-statistics data for 31 large cities published in the 1880 census with no adequate accompanying age data. Massachusetts, Rhode Island, Vermont, Connecticut, and New Jersey also have good mortality statistics by age, but detailed age data were previously available only for Massachusetts. Thus, many of the key mortality data for the late nineteenth century have been unusable for lack of appropriate denominators. Finally, detailed analysis of nuptiality for population subgroups and geographic locations is possible for the first time.

The availability of information on the entire nineteenth-century population opens a variety of new research strategies. Multilevel modeling, for example, requires independent variables tabulated for small geographic units, and such data are generally unavailable for the period before the twentieth century. The new 1880 database allows creation of a wide variety of contextual variables—such as racial or ethnic composition, female labor-force participation, and occupational structure—at any geographic level, including the block, the neighborhood, and the enumeration district.

The 1880 database will also allow analysts to exploit new cartographic technology. Because existing microdata files are samples, when used for small areas they provide insufficient precision for reliable mapping. Geographers are thus forced to rely on complete-count aggregate data that usually provide only basic summary statistics for small areas. The 1880 census database will provide full geographic detail for every individual in the population. There is a large public investment in new historical geographic information systems, including the National Historical Geographic Information System (NHGIS) (see Fitch and Ruggles, pp. 41–51, in Part One of this issue). The 1880 national database will allow scholars to marry these geographic boundary files to population characteristics, thus creating a powerful new analytic tool. Such fine geographic analysis will be especially potent in the analysis of topics such as early suburban development and racial and ethnic residential segregation (see, e.g., Gardner 1998).

Perhaps most important, the creation of a complete-count database for 1880 gives researchers unprecedented opportunity to study small population subgroups. Analysts of particular communities can extract their entire population; as shown in table 1, this represents a large improvement over the existing 1 percent sample data. Other subpopulations of interest are geographically dispersed. Table 2 reports the number of cases for various population subgroups in the 1 percent sample of the 1880 population and the expected number in the new 1880 database. With the sample data, immigration historians were limited to the study of the largest ethnic groups, principally the Germans and Irish (King and Ruggles 1990). The new database allows for study of the first wave of immigrants from such countries as Italy, Poland, China, and Mexico. Labor historians and analysts of professionalization can study a wide range of specific occupational groups. Current analyses of long-run change in marital instability were plagued by the shortfall of cases in the existing 1880 sample; this problem is eliminated by the availability of the comprehensive 1880 database (Ruggles 1997). In each of these examples, a complete enumeration offers a sufficient number of cases to answer basic demographic research questions.

TABLE 1. Comparison of the Number of Persons in the 1% Sample and Complete Database of the 1880 Census, by Selected Subpopulations

Subpopulation	1880 census	
	Current 1% sample	Complete 100% database
Immigrant groups		
Mexico	746	74,400
Norway	2,112	188,658
Ireland	18,757	1,870,830
Netherlands	588	58,654
Italy	426	42,451
Czechoslovakia	929	92,694
Poland	653	65,115
China	1,140	113,733
Occupational groups		
Clergymen	696	69,397
Lawyers and judges	660	65,815
Physicians and surgeons	894	89,182
Midwives	26	2,592
Policemen and detectives	149	14,858
Boot- and shoemakers	1,899	189,381
In Lynn, Mass.	115	11,474
Cotton-mill operatives	1,900	189,450
In Lowell, Mass.	78	7,787
Divorced persons		
Women	588	58,636
Men	326	32,526

TABLE 2. Comparison of the Number of Persons in the 1% Sample and Complete Database of the 1880 Census, by Selected Urban Populations

Urban area	1880 census: Current 1% sample	1880 census: Complete 100% database
Atlanta	409	37,409
Baltimore	3,355	332,313
Boston	3,596	362,839
Chicago	4,918	503,185
Cincinnati	2,629	255,139
Galveston	290	22,248
Los Angeles	not identified	11,183
Milwaukee	1,130	115,587
Minneapolis	382	46,887
New Orleans	2,116	216,090
New York	11,781	1,206,299
Norfolk	262	21,966
Pittsburgh	1,531	156,389
Sacramento	230	21,420
St. Joseph	291	32,431
St. Louis	3,599	350,518
Salt Lake City	198	20,768
San Francisco	2,425	233,959
Washington, D.C.	1,617	147,293

Source Data

The data files created by the LDS are literal transcriptions of the original manuscripts, except for standardized spellings of some birthplaces and occupations and abbreviations of family relationships. Unfortunately, the LDS volunteers did not transcribe all the information on the census form; to save time, they omitted information relating to disability (the condition of being sick, blind, deaf, maimed, idiotic, or insane), literacy, school attendance, months unemployed, whether an individual had married within the census year, and month of birth.[2] As described below, we plan to seek funding to recover this information for a subset of the population. The LDS volunteers did, however, transcribe most of the key variables: county, locality, microfilm reel and page, name, age, sex, race, marital status, relation to head, occupation, birthplace, and parental birthplaces.

In preparation for their genealogical look-up system, the LDS staff converted the raw data into an Oracle database format and carried out a variety of edits. Some of this work was labor-intensive, involving manual examination of millions of cases with invalid entries. These corrections are invaluable and represent many thousands of hours of hand labor. Other aspects of the cleaning process were automated, such as the elimination of duplicate cases. Unfortunately, this cleaning work was flawed, and in the process the LDS lost about 10 percent of the cases. Moreover, the new database inadvertently dropped two key variables: the batch number, needed to identify each case uniquely, and the "new household" flag, which identifies the beginning of each new household.

When these errors were discovered, the LDS staff returned to the raw data and created a new Oracle database that included the missing information. We were presented with two versions of the database: an "original" version, which was nearly complete but had not yet been cleaned, and a "processed" version, which was missing some records and key variables. To recover the lost cases and still take advantage of the LDS editing work, we merged the two versions, a task greatly complicated by the lack of unique identifiers in the processed version.

Our work was also complicated by the fact that the LDS had a publishing schedule; they had requested that we complete our work by January 2001. We began processing the data in fall 1999 with the understanding that we would give highest priority to issues of greatest concern to the LDS. These tasks included restoring missing and corrupted locator information (e.g., microfilm reel and page information, as well as variables used to order cases within a given page, assigning basic geographic information—county, minor civil division (MCD), and incorporated municipality—for the entire database, identifying missing or duplicate records, correcting internal data inconsistencies, and verifying breaks between households.

Editing Locator Information and Geographic Fields

Geographic analysis provided an important tool for handling many of the problems with the 1880 database. The 1880 population volume gives a detailed listing of population totals for MCDs and incorporated municipalities, sorted by state and county (U.S. Census Office 1883). The microfilm reels of the census enumeration forms are organized by enumeration district, county, and state (National Archives 1979). Thus, we could identify errors in microfilm reel number, state, and county by searching for inconsistencies between these fields. When reel and page numbers were missing or corrupted for a group of records but geographic information was available, we could typically make an educated guess about the location of the records, which we confirmed by examining the relevant microfilm.

After correcting reel number, page number, and state and county of residence, we began to deal with missing or inconsistent geographic information at the subcounty level, including information regarding MCDs (often townships, but known as towns, beats, or precincts in some states) and incorporated municipalities (cities, towns, and villages). Because this information is listed at the top of census manuscript pages, the LDS data-entry volunteers did not enter geographic information for each record. Rather, they used a secondary data-entry form for this information, which was supposed to be appended to all records from a given census page at a later processing stage. This process, how-

ever, was inconsistently executed, leaving us with a large number of records with missing or inconsistent information regarding MCDs and incorporated municipalities. Because it was an extensive problem, we found it cost-effective to automate much of this work. If some records for a census lacked subcounty geographic information but other records for the same page had legitimate and consistent identifiers, we automatically filled in the geographic information for the entire page. When the information for the records from a given census page was inconsistent, however, research assistants manually consulted the microfilm to determine the correct information.

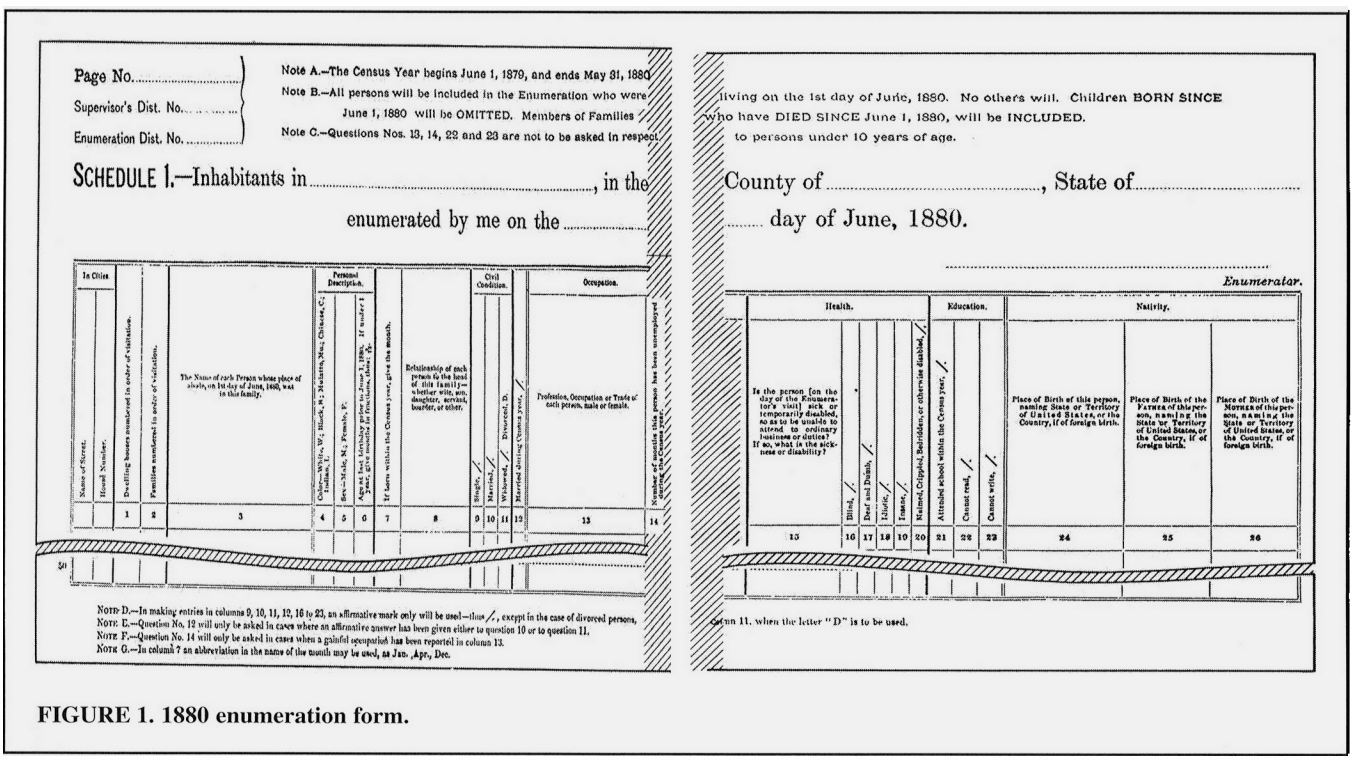
Assigning correct subcounty geographic information was complicated by the layout of the 1880 census form. The top of each page has a series of blank lines preceded by "Inhabitants in," "in the County of" and "State of" (see figure 1). The state and county fields were generally filled in accurately by enumerators, but the field pertaining to inhabitants was a source of problems. Whereas in some states, MCDs and incorporated municipalities are mutually exclusive, in most states, villages and small cities are contained within MCDs. Some enumerators conscientiously recorded both the name of the MCD as well as that of the incorporated municipality in this field, but more typically an enumerator recorded only one or the other. Many incorporated villages and small cities were not listed in this field at all; rather, enumerators indicated these place names in the page margin or at the bottom of the page when they completed the enumeration of a specific village or small city. Also problematic were cases where the MCD and incorporated municipality names were identical: for example, when enumerating residents of Washington Village in Washington Township, an enumerator often recorded "Washington" without indicating whether it was the village or the township.

Distinguishing between an MCD and an incorporated municipality is an important issue for social science data because this information is needed to assign urban or rural status. This distinction was a low priority for the LDS because the primary use of geographic information for their genealogical database is to facilitate searches for specific individuals. For our 2001 LDS deadline, therefore, we focused on filling in MCD information for the entire database.

Since delivering the cleaned data to the LDS, we have resumed the laborious task of identifying villages and small cities. We are also assigning the 1880 published population totals and Federal Information Processing Standards (FIPS) codes to all MCDs and incorporated municipalities. This work has been time consuming, but a complete-count database with accurate geographic information and population totals at the subcounty level will be a valuable resource for social science research.

Missing and Duplicate Records

The identification of missing and duplicate records was an issue of great concern to the LDS, and we have developed three processes to address these problems. First, we


Page No.....
Supervisor's Dist. No.....
Enumeration Dist. No.....

Note A.—The Census Year begins June 1, 1879, and ends May 31, 1880.
Note B.—All persons will be included in the Enumeration who were living on the 1st day of June, 1880. No others will. Children BORN SINCE June 1, 1880 will be OMITTED. Members of Families who have DIED SINCE June 1, 1880, will be INCLUDED.
Note C.—Questions Nos. 13, 14, 22 and 23 are not to be asked in respect to persons under 10 years of age.

SCHEDULE 1.—Inhabitants in....., in the County of....., State of.....
enumerated by me on the ..... day of June, 1880.

.....
Enumerator.

In Cities		Dwelling houses numbered in order of visitation.	Families numbered in order of visitation.	The Name of each Person whose place of abode, on 1st day of June, 1880, was in this family.	Personal Description.				Relationship of each person to the head of this family—whether wife, son, daughter, servant, boarder, or other.	Civil Condition.				Occupation.		Health.						Education.			Nativity.		
Street.	House Number.				Color—White, W.; Black, B.; Mulatto, Mu.; Chinese, C.; Indian, I.	Sex—Male, M.; Female, F.	Age at last birthday prior to June 1, 1880. If under 1 year, give months.	If born within the Census year, give the month.		Single.	Married.	Widowed.	Divorced.	Profession, Occupation or Trade of each person, male or female.	Number of months this person has been unemployed during the Census year.	Is the person (on the day of the Enumerator's visit) sick or temporarily disabled, so as to be unable to attend to ordinary business or duties? If so, what is the sickness or disability?	Blind.	Deaf and Dumb.	Idiotic.	Insane.	Maimed, Crippled, Bedridden, or otherwise disabled.	Attended school within the Census year.	Cannot read.	Cannot write.	Place of Birth of this person, naming State or Territory of United States, or the Country, if of foreign birth.	Place of Birth of the Father of this person, naming the State or Territory of United States, or the Country, if of foreign birth.	Place of Birth of the Mother of this person, naming the State or Territory of United States, or the Country, if of foreign birth.
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26

Note D.—In making entries in columns 9, 10, 11, 12, 16 to 23, an affirmative mark only will be used—thus /, except in the case of divorced persons, Note E.—Question No. 12 will only be asked in cases where an affirmative answer has been given either to question 10 or to question 11.
Note F.—Question No. 14 will only be asked in cases when a gainful occupation has been reported in column 13.
Note G.—In column 7 an abbreviation in the name of the month may be used, as Jan., Apr., Dec.

Column 11, when the letter "D" is to be used.

FIGURE 1. 1880 enumeration form.

compared the number of records for specific counties with the published Census Bureau population totals. Research assistants examined the microfilm and database when the discrepancy between the number of records and the population totals exceeded 100. Typically, this examination disclosed that the database was missing individuals from a specific range of manuscript pages. We then notified the LDS staff, who searched their data archive for the missing cases. In the majority of cases, the LDS was able to recover the missing data. When they could not do so, our research assistants entered the missing records.

The second procedure for identifying missing cases involved searching for missing and incomplete pages. From the database, we constructed a table listing the number of records for each manuscript page on each microfilm reel. We used this table to identify missing pages in the database by flagging any breaks in the sequence of pages. We were able to identify partially missing pages and duplicate pages by examining those pages for which the total number of records was less than 90 or greater than 100.[3]

These procedures were successful in restoring or inserting large numbers of missing records. When we began a more systematic examination of households, however, research assistants began to notice random missing records. We suspect the cause was database corruption during data conversion processes, but it was difficult to devise a method for identifying these deleted records. A visual examination of pages with fewer than 100 records was impractical; in 1880, many pages include blank lines, and we did not have sufficient resources to examine every page manually. Fortunately, we were able to construct a third diagnostic procedure capitalizing on a unique feature of the LDS data-entry process. As part of their administrative procedures, the LDS assigned each record a sequential entry number. Thus, the first record entered by each data-entry operator was assigned an entry number of "1," and the entry number increased sequentially with each additional record. The size of entry number sequences for the entire database ranges from approximately 1,000 to 5,000 cases. Any break in the entry number sequence—for example, if the entry numbers for two consecutive records were "482" and "484," respectively—usually indicates a dropped record. In such cases, a research assistant examined the microfilm to determine if a case was missing. This process enabled us to restore approximately 30,000 records that otherwise would have been difficult to identify.

At the time of our preliminary release, the database contained 50,480,213 individuals compared with the official population in 1880 of 50,155,783.[4] That discrepancy is partly a consequence of double counting St. Louis; the city was enumerated twice in 1880, and both enumerations are included in the database. Subtracting individuals from the first enumeration of St. Louis leaves a total of 50,158,577. Thus, we have a surplus of 2,794 records. Given the condition of the database when we began the project, we are elated to be within 3,000 cases of the Census Bureau's published population total for 1880.

The figure of 2,794 is our net error and reflects both extra records and missing records. The extra records result from the inclusion of persons who were crossed off the original manuscripts. Some marks were made by enumerators or Census Bureau clerks who determined that the individual had been enumerated elsewhere. Other individuals are explicitly listed as deceased at the time of the census, a determination made by clerks who compared the names with those on mortality schedules. LDS data-entry volunteers entered these individuals into the database, and we have no plans to remove them. Nonetheless, we will construct a variable indicating their status. If we excluded crossed-off cases, the case count in the database would probably be lower than the published population total. We are continuing to identify missing records by closely examining enumeration districts where we had previously found missing records. We also plan further examination of discrepancies between the database and published population totals for MCDs and incorporated municipalities.

Household, Family, and Consistency Edits

A significant problem with the 1880 database stemmed from the LDS procedure for designating breaks between households. The 1880 census had separate columns for dwelling and family numbers, and enumerators were instructed to sequentially number both units within their districts. A census family was defined as an individual or group of individuals (including unrelated individuals such as boarders and servants) living together in the same dwelling place. Thus, a census family closely resembles the current census definition of a household. Two or more families could reside in a single structure, provided they occupied separate quarters and their housekeeping was separate. Wherever such multifamily dwellings occur, the count of dwellings will be lower than the count of families.[5]

The LDS data-entry volunteers did not enter dwelling or family numbers; rather, they indicated household breaks by inserting a "P" in a specified field for the first person in a household. We determined that a significant number of household breaks were actually breaks between dwellings. We identified these cases by searching for households containing multiple heads or multiple wives and then inserting a new household indicator when appropriate.

We encountered another problem in that LDS volunteers occasionally disregarded the original enumeration of households in an attempt to delineate nuclear families. In such cases, unrelated individuals such as boarders and servants were given a new household designator, despite being clearly listed as coresident, unrelated individuals on the microfilm. We corrected this problem by examining instances where the database contained three consecutive household heads. If we determined that the household breaks were

incorrect, a research assistant examined the remainder of the enumeration district manually, correcting any additional household errors.

We combined work on fixing household breaks with other logical edit checks. A common inconsistency involved conflicting information for sex and relationship to head, such as an individual identified as "male" with a relationship of "daughter." We automated some of this work by creating gendered given-name lists. Thus, if the male daughter also had a first name of Abigail, we changed the sex to female. If the name was ambiguous, however, a research assistant consulted the microfilm to determine the correct information.

We carried out a variety of other data-consistency checks. These included nonstandard responses for sex, race, and marital status; household heads under the age of 16; wives with a marital status of "single"; implausible age gaps between spouses and between parents and children; and sons and unmarried daughters with a surname different from the household head's. Our checking program flagged approximately 400 errors per microfilm reel; each error required a research assistant to examine the microfilm and correct erroneous information as needed. Some individual error flags required correcting information for multiple individuals in a given household. Ultimately, we devoted over ten thousand hours to the data-checking process.

During the past year, we have continued to run diagnostic programs designed to identify potential errors. Most of this work has continued to focus on problematic household breaks and erroneous relationships to household heads. For example, we flagged all coresident individuals with an "X" in the relationship-to-head field (indicating an unrelated individual) who also had the same surname as the household head. Many of these individuals actually had a missing relationship to head on the microfilm; apparently some data-entry volunteers used an "X" to indicate both unrelated individuals and those with missing information. We edited the relationship-to-head field for those individuals to indicate that the field was actually blank on the microfilm.

Numeric Coding and Constructed Variables

We are constructing numeric codes for all variables that are compatible with the coding designs of the Integrated Public Use Microdata Series (IPUMS) and the North Atlantic Population Project (NAPP). This work builds on data dictionaries we developed for the 1880 sample and related data projects, but the large scale of the database means that numeric coding remains a major undertaking, especially for birthplaces and occupations.

Although we supplied birthplace codes for our preliminary release of October 2001, we are currently refining the coding process for ambiguous entries. For example, approximately 13,000 records have a birthplace of "CHI." Most of these individuals were born in China, but a significant number were born in Chicago, and some were born in Chihuahua (Mexico). As a result, we have begun examining birthplace information in combination with other variables such as surname and place of residence to help resolve these ambiguous cases.

The occupational-data dictionary requires coding of more than 500,000 unique occupational strings. When the dictionary is complete, all individuals in the labor force will have their occupations coded according to three different occupational coding schemes: the 1880 Census Bureau occupational coding system, the 1950 Census Bureau occupational coding system, and a modified version of the Historical International Standard Classification of Occupations (HISCO) coding system (see Roberts et al., article on occupational classification in NAPP in Part Two of this issue). In addition, the final release of the database will contain the text strings for most variables, which will allow researchers to devise modified coding schemes to accommodate their specific research agendas.

The database will provide IPUMS-compatible constructed variables. All individuals will have pointer variables allowing users to easily link husbands to their wives and children to their parents. We will construct family-level variables listing the number of own-family members, siblings, children, and children under the age of 5. We will also include the constructed occupational and geographic indicators used in the IPUMS, including socioeconomic index (SEI), occupational score, labor-force status, incorporation status, metropolitan status, metropolitan area, metropolitan district, and state economic area.

The final version of the database will also employ allocation procedures for missing data. This process involves searching the data file for a "donor" record that shares key characteristics with each missing, illegible, or inconsistent case. For example, if the sex is missing or illegible and cannot be logically inferred from other characteristics, the sex of the most proximate person record with the same household type and size, race, age, marital status, and relation to household head is allocated. We will provide data-quality flags that will allow users to identify allocated data. A detailed description of these procedures will be integrated with documentation currently available at the IPUMS Web site.

Current Status and Future Plans

In October 2001, we released a preliminary version of the database that reflects all the data cleaning we performed for the initial LDS release, along with a few enhancements. We inserted supervisor and enumeration district information for the entire database. We also examined all households containing 10 or more individuals unrelated to the household head so as to determine if the household was part of an

institution and whether the institution's name was listed on the census manuscript. In addition, we added numeric codes for birthplace and parental birthplaces based on a partially coded birthplace dictionary. Our next goal is to release a numerically coded complete-count database compatible with the existing 1 percent microdata samples. We anticipate completing this work by September 2003.

The greatest limitation of the complete-count database for 1880 is the omission of key variables. As noted above, the LDS did not enter information on sicknesses on the day of enumeration, disability (the condition of being sick, deaf, maimed, idiotic, or insane), literacy, school attendance, months unemployed, whether an individual had married within the census year, and month of birth. In addition, the LDS did not enter information on household relationship for persons who were unrelated to the household head, so, for example, it is impossible to distinguish boarders from servants. It would be prohibitively expensive to add the missing information to the entire census, so we plan to propose creating a 10 percent sample with complete information. We will thus extend the series of planned high-density samples for 1900, 1930, and 1960 described elsewhere (in the introduction by Ruggles, pp. 5–8 in Part One of this issue, and Ruggles et al.—IPUMS Redesign—pp. 9–19 in Part One of this issue).

As part of the same project, we plan a pilot study to link individuals in the 1880 sample to adjacent census years. Perhaps the greatest limitation of the existing microdata samples is that they are cross-sectional snapshots and do not allow one to trace individuals across time. This problem will be greatly alleviated by the new 1880 database. The MPC has prepared 1-in-100 samples from 1850, 1860, 1870, 1880, 1900, 1910, and 1920. Each of these samples is independent, so they cannot be linked together to provide two observations for the same individual. It will be possible, however, to link each of these samples to the new 1880 database. We can therefore construct six linked samples, covering 1850–1880, 1860–1880, 1870–1880, 1880–1900, 1880–1910, and 1880–1920.

Past local studies linking individuals across census years lost a high percentage of cases because of migration (e.g., see Katz 1975, Knights 1991; Thernstrom 1964). In 1987, Thomas Pullum and Avery Guest created a national linked "panel" of two cohorts of men in the 1880 and 1900 censuses. Between these two census years, they managed to link 4,014 individuals out of a sample of 10,252, for a linkage rate of 39.4 percent (Guest 1987). More recently, Joseph Ferrie (1996) linked a nationally representative subset of the 1850 public use microdata sample to the 1860 manuscript census. Ferrie limited his study to persons with uncommon surnames but still achieved only a 19.3 percent linkage rate for a total of 4,938 linked cases.

The availability of a high-quality census file including the entire 1880 population will allow far more sophisticated matching than has previously been possible. The Pullum/Guest and Ferrie samples had to rely on the state Soundex name indexes to locate individuals in the 1900 census. Interstate migrants were for the most part lost. The process was labor-intensive, extremely expensive, and involved a large potential for human error. Using the new 1880 database, however, the entire country can be searched using such characteristics as age, sex, birthplace, birthplace of mother, and birthplace of father as well as name. We therefore expect that the 1880 database will not only allow a far higher rate of matches than have previous studies but will also be able to provide samples significantly larger and more representative than the Pullum/Guest and the Ferrie studies. In the end, a series of nationally representative linked samples for the nineteenth century may prove to be the most substantial contribution of the 1880 complete population database.

NOTES

We thank the Church of Jesus Christ of Latter-day Saints (LDS) for their efforts to create the 1880 database and their willingness to allow us to make a version available for academic research. We also thank the National Institutes of Health (HD 39327) and the National Science Foundation (SES 9910961) for funding to complete this project.

1. The preliminary version of the 1880 U.S. population database is available at www.nappdata.org. Eventually, the U.S. database will be integrated with data from Great Britain, Canada, Norway, and Iceland to form the North Atlantic Population Project (NAPP). Two articles by Roberts et al. (NAPP Overview and Occupational Classification in NAPP) appear in Part Two of this issue.
2. In addition, the LDS did not enter street-address information and family and dwelling numbers. Another peculiarity is that they did not record the actual relationship to head for individuals unrelated to the household head (e.g., boarders, lodgers, and servants). Rather, unrelated individuals were denoted with an "X" in the relationship-to-head field.
3. Although each manuscript page in 1880 has 50 lines, in the process of microfilming, the national data archives assigned a stamped page number consisting of two sides; for example, page 25 on a given reel has an "A" side and a "B" side. Thus, each page has 100 lines with a theoretical maximum of 100 records per page. The minimum number of records per page is more problematic because many enumerators left partially blank pages at the end of an enumeration of a specific minor civil division (MCD) or village. In fact, some enumerators left a blank line between households.
4. In addition to the states, the database contains individuals from most territories that eventually became the continental United States. The sole exception is Indian Territory, which mainly became present-day Oklahoma.
5. Although more common today, multifamily dwellings in 1880 were not a rare occurrence, especially in larger cities.

REFERENCES

Eckler, A. R. 1972. *The Bureau of the Census.* New York: Praeger.

Ferrie, J. P. 1996. A new sample of males linked from the public use microdata sample of the 1850 U.S. federal census of population to the 1860 U.S. federal census manuscripts. *Historical Methods* 29: 141–56.

Gardner, T. 1998. *The rise of the American suburb, 1850–1950.* Ph.D. diss., University of Minnesota.

Guest, A. 1987. Notes from the national panel study: Linkage and migration in the late nineteenth century. *Historical Methods* 20: 63–77.

Katz, M. B. 1975. *The people of Hamilton, Canada West: Family and class in a mid-nineteenth-century city.* Cambridge: Harvard University Press.

King, M. L., and S. Ruggles. 1990. American immigration, fertility differentials, and the ideology of race suicide at the turn of the century. *Journal of Interdisciplinary History* 20: 347–69.

Knights, P. R. 1991. *Yankee destinies: The lives of ordinary nineteenth-century Bostonians.* Chapel Hill: University of North Carolina Press.
Leet, D. R. 1977. Interrelations of population density, urbanization, literacy, and fertility. *Explorations in Economic History* 14(4): 388–410.
Magnuson, D. L. 1995. *The making of a modern census: The United States census of population, 1790–1950.* Ph.D. diss, University of Minnesota.
National Archives. 1979. *Federal population censuses, 1790–1890: A catalog of microfilm copies of the schedules.* Washington, D.C.: National Archives Trust Fund Board.
Ruggles, S. 1997. The rise of divorce and separation in the United States, 1880–1980. *Demography* 34: 455–66.
Ruggles, S., and M. Sobek. 1998. *IPUMS-98: Integrated Public Use Microdata Series.* 5 vols. Minneapolis: Historical Census Project, University of Minnesota.
Schapiro, M. O. 1985. A general dynamic model of the 19th century U.S. population change. *Economic Modelling* 2(4): 347–56.
Scott, A. H. 1968. *Census U.S.A.: Fact finding for the American people, 1790–1970.* New York: Seabury Press.
Thernstrom, S. 1964. *Poverty and progress: Social mobility in a nineteenth-century city.* Cambridge: Harvard University Press.
U.S. Census Bureau. 1975. *Historical statistics of the United States: Colonial times to 1970.* Series B67–98. Washington, D.C.: GPO.
U.S. Census Office. 1882. Report of the Superintendent of the Census, November 1, 1881. Washington, D.C.: GPO.
_____. 1883. The tenth census of the United States: 1880. *Statistics of population.* Washington, D.C.: GPO).
_____. 1895. The eleventh census of the United States: 1890. *Population, Part 1.* Washington, D.C.: GPO.
Walker, F. A. 1888. The eleventh census of the United States. *Quarterly Journal of Economics* 2: 136–61.
Wright, C. D., and W. Hunt. 1900. *The history and growth of the United States census.* Washington, D.C.: GPO.